Data Sharing – Best Practice und Perspektiven:

# Archiv für Gesprochenes Deutsch

Thomas Schmidt & Arnulf Deppermann

Institut für Deutsche Sprache, Mannheim

# Gliederung

1. Hintergrund
   - IDS und Archiv für gesprochenes Deutsch (AGD)
   - Datenbank für gesprochenes Deutsch (DGD)
2. Elemente einer Best Practice für Gesprächsdaten
   - Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)
3. Perspektiven
   - Forschungsinfrastrukturen

# Institut für Deutsche Sprache (IDS)

- Mitglied der Leibniz-Gemeinschaft
- Einziges außeruniversitäres Forschungsinstitut zur Dokumentation und Erforschung der deutschen Sprache der Gegenwart
- Infrastrukturknotenpunkt der Linguistik: Korpora, korpustechnologische Instrumente

# IDS und qualitative Sozialforschung

- Forschung der Abtlg. „Pragmatik“ des IDS: Untersuchung der verbalen Interaktion in Alltag und Institutionen (Konversationsanalyse, multimodale Videoanalyse)

→Erstellung, Archivierung und öffentliche Bereitstellung von Gesprächskorpora und variationslinguistischen Korpora

# Archiv für gesprochenes Deutsch

- hervorgegangen aus dem Deutschen Spracharchiv (seit 1957), am IDS seit 1979
- Seit 2004 als AGD in der Abteilung Pragmatik des IDS
- fortgeschrittene Digitalisierung

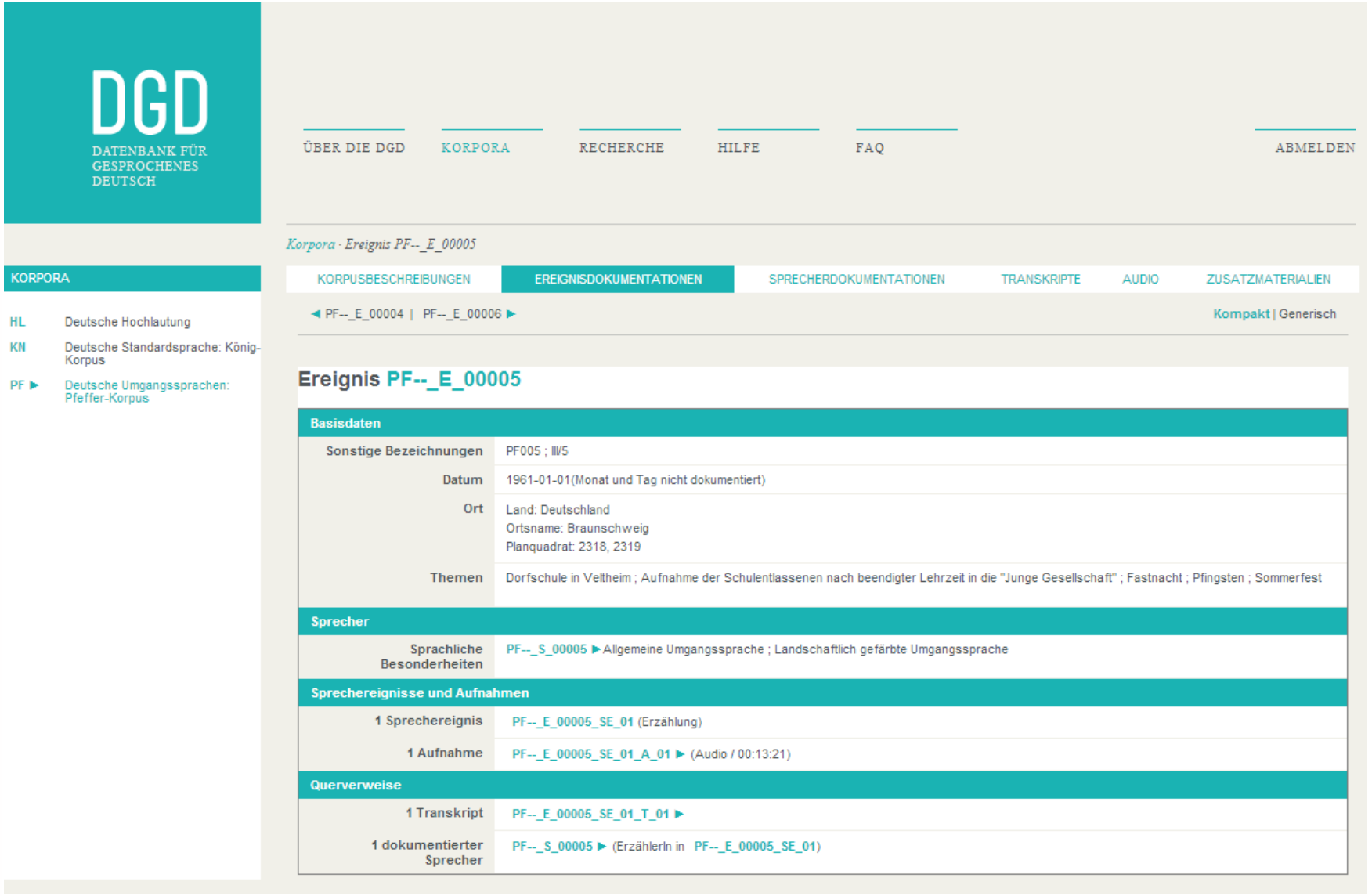
- Datenbestände
  - Übernahme und Aufbereitung aus abgeschlossenen Projekten (intern und extern)
  - Varietäten-Korpora (deutsche Dialekte, auslandsdeutsche Varietäten)
  - Gesprächskorpora (z.B. Stadtsprache Mannheim, Berliner Wendekorpus)
  - über 6000 h Tonaufnahmen, über 6 Mio transkribierte Wort-Token

- Nutzer
  - Dialektologie
  - Gesprächsforschung
  - Lexikographie, Grammatik, Sprachwandel, Soziolinguistik, Sprachlehre…
  - Kulturwissenschaft, Qualitative Sozialforschung, …
  - Ausstellungen, Kunst,  Vereine, …

- Distribution der Daten im individuellen Service und über die DGD

# Datenbank für Gesprochenes Deutsch

- DGD1: seit 1997
- Browsen, Recherche, Download von Metadaten, Aufnahmen, Transkripten
- Substantieller Teilbestand des AGD
- Aktuell: DGD2 als modernisierte Version
  - Transformation der Datenbestände in aktuelle Standards (XML), Qualitätskontrolle
  - Erweiterung der Datenbestände um FOLK
  - Erweiterung der Funktionalität: Korpuslinguistische Recherche und Analyse




AGD

Thomas Schmidt / Arnulf Deppermann
Archiv für gesprochenes Deutsch

INSTITUT FÜR DEUTSCHE SPRACHE

*Korpora* · *Transkript FOLK_E_00047_SE_01_T_02*

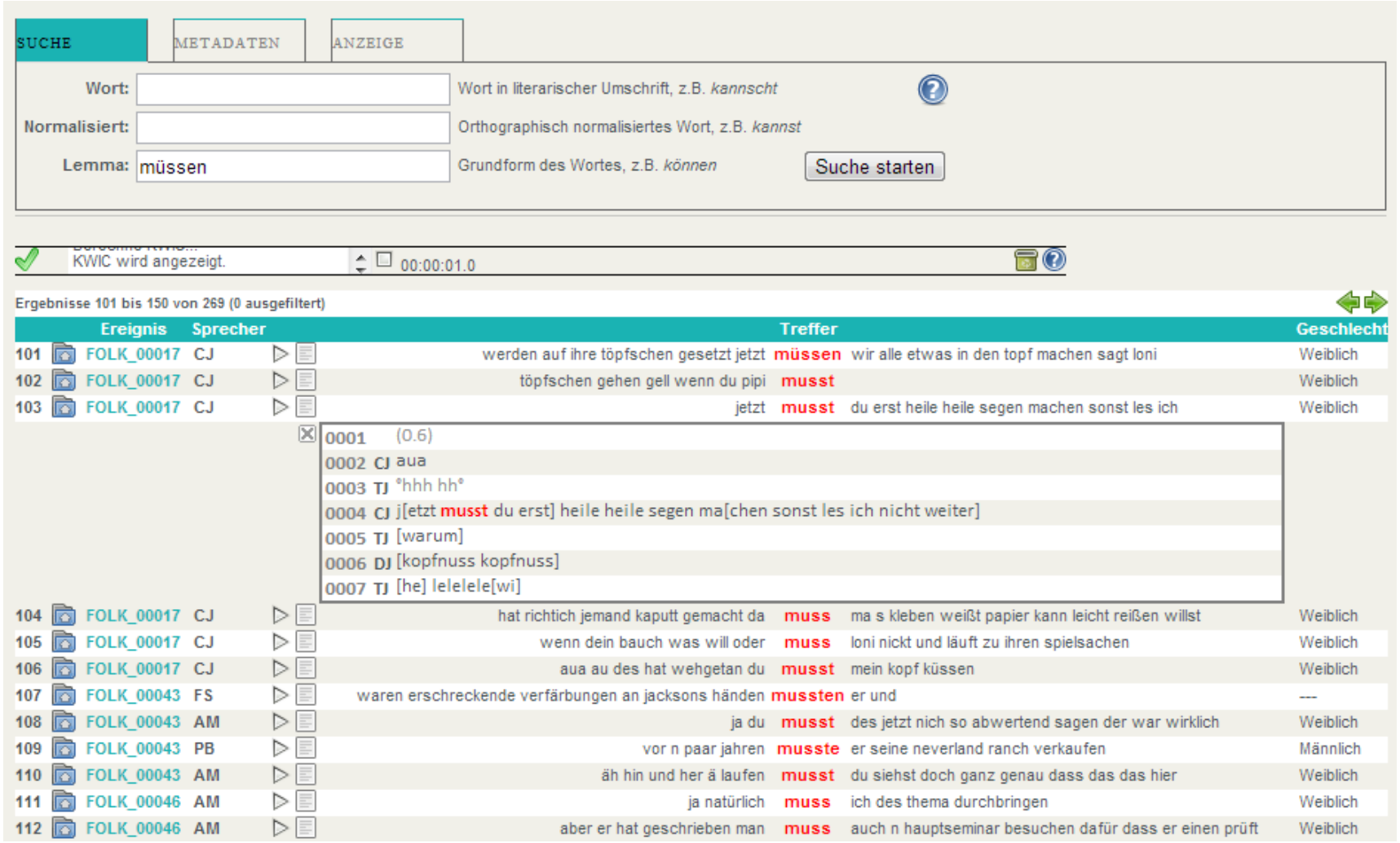
KORPUSBESCHREIBUNGEN | EREIGNISDOKUMENTATIONEN | SPRECHERDOKUMENTATIONEN | **TRANSKRIPTE** | AUDIO | ZUSATZMATERIALIEN

◄ FOLK_E_00047_SE_01_T_01 | FOLK_E_00050_SE_01_T_01 ►

00:32:29.87

**Doppelklick** auf eine Stelle im Transkript zum Starten der alignierten Aufnahme (15-Sekunden Ausschnitt)
Klick auf den **Stop-Button** zum Anhalten der alignierten Aufnahme

| | | | |
|---|---|---|---|
| {32:49} | 018 | AM | ja ich |
| {32:49} | 019 | | (0.2) |
| {32:49} | 020 | PB | des heißt du sitzt da voll klimatisiert drin und kriegst diesen ganzen staub nich ab |
| {32:54} | 021 | | (1.23) |
| {32:55} | 022 | PB | oder es gibt da diese |
| {32:56} | 023 | | (0.58) |
| {32:56} | 024 | PB | kleineren geräde |
| {32:57} | 025 | | (0.2) |
| {32:58} | 026 | PB | °h wo du so keine (.) äh |
| {32:59} | 027 | | (1.89) |
| {33:01} | 028 | PB | kabine hast |
| {33:02} | 029 | | (0.2) |
| {33:02} | 030 | PB | bei (.) achteneunzig grad im schatten |
| {33:05} | 031 | | (2.36) |
| {33:07} | 032 | PB | da mähdrescher fahren musst und den ganzen ähm |


SUCHE
METADATEN
ANZEIGE
Wort:
Wort in literarischer Umschrift, z.B. kannscht
Normalisiert:
Orthographisch normalisiertes Wort, z.B. kannst
Lemma:
müssen
Grundform des Wortes, z.B. können
Suche starten
KWIC wird angezeigt.
00:00:01.0
Ergebnisse 101 bis 150 von 269 (0 ausgefiltert)
Ereignis Sprecher Treffer Geschlecht
101 FOLK_00017 CJ werden auf ihre töpfschen gesetzt jetzt müssen wir alle etwas in den topf machen sagt loni Weiblich
102 FOLK_00017 CJ töpfschen gehen gell wenn du pipi musst Weiblich
103 FOLK_00017 CJ jetzt musst du erst heile heile segen machen sonst les ich Weiblich
0001 (0.6)
0002 CJ aua
0003 TJ °hhh hh°
0004 CJ j[etzt musst du erst] heile heile segen ma[chen sonst les ich nicht weiter]
0005 TJ [warum]
0006 DJ [kopfnuss kopfnuss]
0007 TJ [he] lelelele[wi]
104 FOLK_00017 CJ hat richtich jemand kaputt gemacht da muss ma s kleben weißt papier kann leicht reißen willst Weiblich
105 FOLK_00017 CJ wenn dein bauch was will oder muss loni nickt und läuft zu ihren spielsachen Weiblich
106 FOLK_00017 CJ aua au des hat wehgetan du musst mein kopf küssen Weiblich
107 FOLK_00043 FS waren erschreckende verfärbungen an jacksons händen mussten er und ---
108 FOLK_00043 AM ja du musst des jetzt nich so abwertend sagen der war wirklich Weiblich
109 FOLK_00043 PB vor n paar jahren musste er seine neverland ranch verkaufen Männlich
110 FOLK_00043 AM äh hin und her ä laufen musst du siehst doch ganz genau dass das das hier Weiblich
111 FOLK_00046 AM ja natürlich muss ich des thema durchbringen Weiblich
112 FOLK_00046 AM aber er hat geschrieben man muss auch n hauptseminar besuchen dafür dass er einen prüft Weiblich

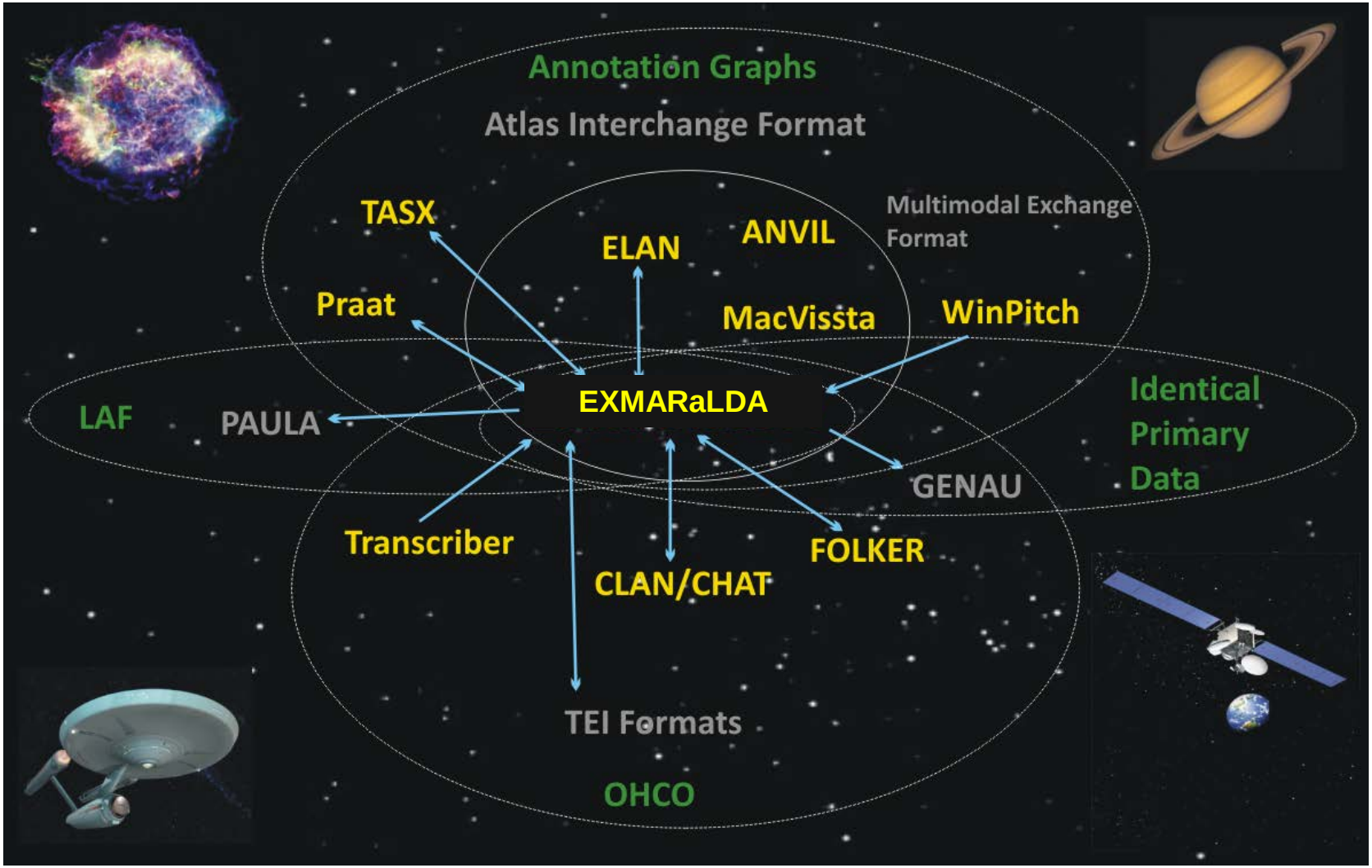
# Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)

- seit 2008 am IDS
- Aufnahmen und Transkriptionen von
  - Privaten Gesprächen
  - Gesprächen in Institutionen (z.B. Unterricht, Prüfungen)
  - Gesprächen im öffentlichen Raum
  - Massenmedial vermittelten Gesprächen (z.B. Schlichtung Stuttgart 21)
- Eigenaufnahmen + Datenspenden
- Ziel: großes, ausgewogenes, computergestützte auswertbares, wissenschafts-öffentlich verfügbares Gesprächskorpus des Deutschen
  Fernziel: Repräsentation des „kommunikativen Haushalts“ (Luckmann) sozialer Interaktionsformen im Deutschen
- Derzeit: ca. 70h, ca. 800.000 transkribierte Wort-Tokens

# Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)

- Entwicklung, Dokumentation, Etablierung einer Best Practice für Gesprächskorpora
  - Datenmodelle und Datenformate
  - Transkription
  - (Metadaten-)Dokumentation
  - Annotationen
  - Software-Tools
  - Rechtliche Aspekte (Datenschutz, Urheberschutz)
  - Publikation

# Datenmodelle und Datenformate



# Datenmodelle und Datenformate

- Modelle: Annotationsgraphen
  → ELAN, Praat, EXMARaLDA, ANVIL, Transcriber, FOLKER, … ~~(F4, inqScribe, Transana)~~
- Formate: Unicode / XML / TEI
- praktikable Interoperabilität, (noch) keine Standardisierung
- **in FOLK: FOLKER-XML-Format**
  - Import und Export z.B. von/nach EXMARaLDA, ELAN, Praat, TEI

# Transkription

- Gesprächsanalytisches Transkriptionssystem (GAT), Selting et al. 1998 und 2009
- Etabliert, weit verbreitet
- (weitere) Spezifizierungen für computergestützte Verarbeitung wünschenswert
- Modularer Aufbau: Minimaltranskript, Basistranskript, Feintranskript
  - wachsender Grad der Abhängigkeit von Theorien und Forschungsinteressen
  - wachsender Aufwand für die Erstellung
  - abnehmende Reliabilität
- **in FOLK: cGAT-Minimaltranskript mit FOLKER**

# GAT-Minimaltranskript

{00:06} 0002 **LB** so ich darf euch begrüßen heut (.) zur unterrichtsstunde wir haben ja zuletzt äh (.) über
{00:11} 0003 (1.03)
{00:12} 0004 **LB** die prüfung
{00:13} 0005 (0.45)
{00:13} 0006 **LB** vom
{00:14} 0007 (0.29)
{00:14} 0008 **LB** sekundärbereich gesprochen
{00:16} 0009 (0.47)
{00:16} 0010 **LB** da müsse mer draußen noch entprechend die (.) tests durchführen
{00:20} 0011 (0.85)
{00:21} 0012 **LB** äh (.) theoretisch ham_mer_s so weit abgeschlossen heut gehen mer einen schritt weidda
{00:24} 0013 (.) ((schmatzt))
{00:25} 0014 °hh
{00:25} 0015 äh
{00:26} 0016 (0.22)
{00:26} 0017 **LB** nennen se mir einfach mal (.) zwei unterschiedliche möglischkeiten
{00:30} 0018 (1.54)
{00:32} 0019 **LB** den (.) primärstrom zu unterbreschen
{00:34} 0020 (7.31)
{00:42} 0021 **LB** ja herr günther
{00:42} 0022 (0.41)
{00:43} 0023 **RG** einma mit unterbrecherkontakt mechanisch oder halt äh elektronisch über steuergerät dann
{00:47} 0024 **LB** gut okay

# GAT-Minimaltranskript

{00:06} 0002 **LB** so ich darf euch begrüßen heut (.) zur unterrichtsstunde wir haben ja zuletzt äh (.) über
{00:11} 0003 (1.03)
{00:12} 0004 **LB** die prüfung
{00:13} 0005 (0.45)
{00:13} 0006 **LB** vom
{00:14} 0007 (0.29)
{00:14} 0008 **LB** sekundärbereich gesprochen
{00:16} 0009 (0.47)
{00:16} 0010 **LB** da müsse mer draußen noch entprechend die (.) tests durchführen
{00:20} 0011 (0.85)
{00:21} 0012 **LB** äh (.) theoretisch ham_mer_s so weit abgeschlossen heut gehen mer einen schritt weidda
{00:24} 0013 (.) ((schmatzt))
{00:25} 0014 °hh
{00:25} 0015 äh
{00:26} 0016 (0.22)
{00:26} 0017 **LB** nennen se mir einfach mal (.) zwei unterschiedliche möglischkeiten
{00:30} 0018 (1.54)
{00:32} 0019 **LB** den (.) primärstrom zu unterbreschen
{00:34} 0020 (7.31)
{00:42} 0021 **LB** ja herr günther
{00:42} 0022 (0.41)
{00:43} 0023 **RG** einma mit unterbrecherkontakt mechanisch oder halt äh elektronisch über steuergerät dann
{00:47} 0024 **LB** gut okay

# (Metadaten-)Dokumentation

- Daten zu Gesprächsumständen, soziobiographischen Sprechereigenschaften, Aufnahmebedingungen
  - (teilweise) korpusabhängig → Katalogdaten / Korpusdesigndaten / Organisatorische Metadaten
  - Teilweise betroffen von Datenschutzbestimmungen
- Einigkeit bezüglich allgemeiner Struktur (Modell)
- Verschiedene Vokabularien, verschiedene Datenformate
- Gemeinsamer Überbau: CMDI-Metadata, ISO-CAT-Registry
- **in FOLK: DGD-Metadatenmodell, FOLK-spezifische Ausgestaltung**

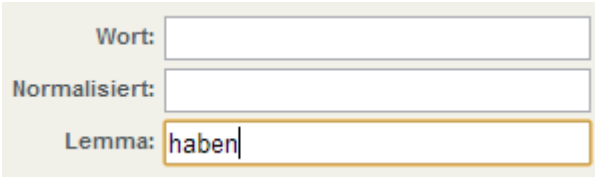
# (Metadaten-)Dokumentation

- **Ereignis: FOLK_E_00001**
  - **Basisdaten:**
    - **Sonstige_Bezeichnungen:** FOLK_BERU_01_A01
    - **Beschreibung:** Unterrichtsstunde in der Berufsschule
    - **Ort:**
      - **Land:** Deutschland
      - **Region:** Rheinfränkische Sprachregion
      - **Kreis:** Anonym
      - **Ortsname:** Anonym
      - **Einwohnerzahl:** 0
      - **Ortsteil:** Nicht dokumentiert
      - **Ortsbeschreibung:** Nicht vorhanden
    - **Institution:** Berufsbildende Schule
    - **Räumlichkeiten:** Klassenzimmer
    - **Datum:**
      - **YYYY-MM-DD:** 2009-03-04
    - **Dauer:** 00:59:45
    - **Zeitraum:** Nicht dokumentiert
    - **Aufnahmebedingungen:** Störungen: teilweise Straßenlärm und Vogelgezwitscher
  - **Projekt:**
    - **Attribut:** Titel="Forschungs- und Lehrkorpus gesprochenes Deutsch"
    - **Personal:**
      - **An_E_teilnehmende_Forscher:** Anonym
      - **An_E_teilnehmende_Techniker:** Nicht vorhanden
  - **Quellaufnahme:**
    - **Attribut:** Kennung="FOLK_E_00001_A_01"
    - **Basisdaten:**
      - **Sonstige_Bezeichnungen:** FOLK_BERU_01_A01
      - **Typ:** Audio
      - **Dauer:** 00:59:45.000000

# Annotationen

- literarische Umschrift → orthographische Normalisierung
  → Lemmatisierung → Part-Of-Speech-Tagging

| theoretisch | ham | mer | s | abgeschlossen | heut | gehn | mer | einen | schritt | weiter |
|---|---|---|---|---|---|---|---|---|---|---|
| | haben | wir | es | | heute | gehen | wir | | Schritt | |
| *theoretisch* | *haben* | *wir* | *es* | *abschließen* | *heute* | *gehen* | *wir* | *ein* | *Schritt* | *weiter* |
| **ADJ** | **V** | **PRO** | **PRO** | **V** | **ADV** | **V** | **PRO** | **ART** | **N** | **ADV** |




| Treffer | | |
|---|---|---|
| irgendwo isch u gees | **hawwen** | se des |
| | **hawwen** | se null und minus |
| sondern isch | **hätt** | jetzt gern zwei begriffe welsche begriffe hätt isch gern |
| jetzt gern zwei begriffe welsche begriffe | **hätt** | isch gern |
| isch | **hab** | keine verbindung |
| das machen isch eigentlich egal er | **hat** | natürlich rescht er sacht warum muss ich jetzt gegen |
| klar wenn sie jetzt keine spannung | **haben** | |
| | **hab** | isch dort keine spannung was wird dann ihr problem |
| des isch no so n tüpp | **ham** | mer geschtern gehabt so n tipp für die |
| noch wenn isch escht ä problem | **hab** | wo isch nischt mehr weiterkomm |
| sehr schön | **ham** | sie gut beantwortet ja also des müsste des gleiche |
| vor allem dann interessant leut mer | **haben** | geschtern so en fall gehabt |
| mer haben geschtern so en fall | **gehabt** | |

# Annotationen

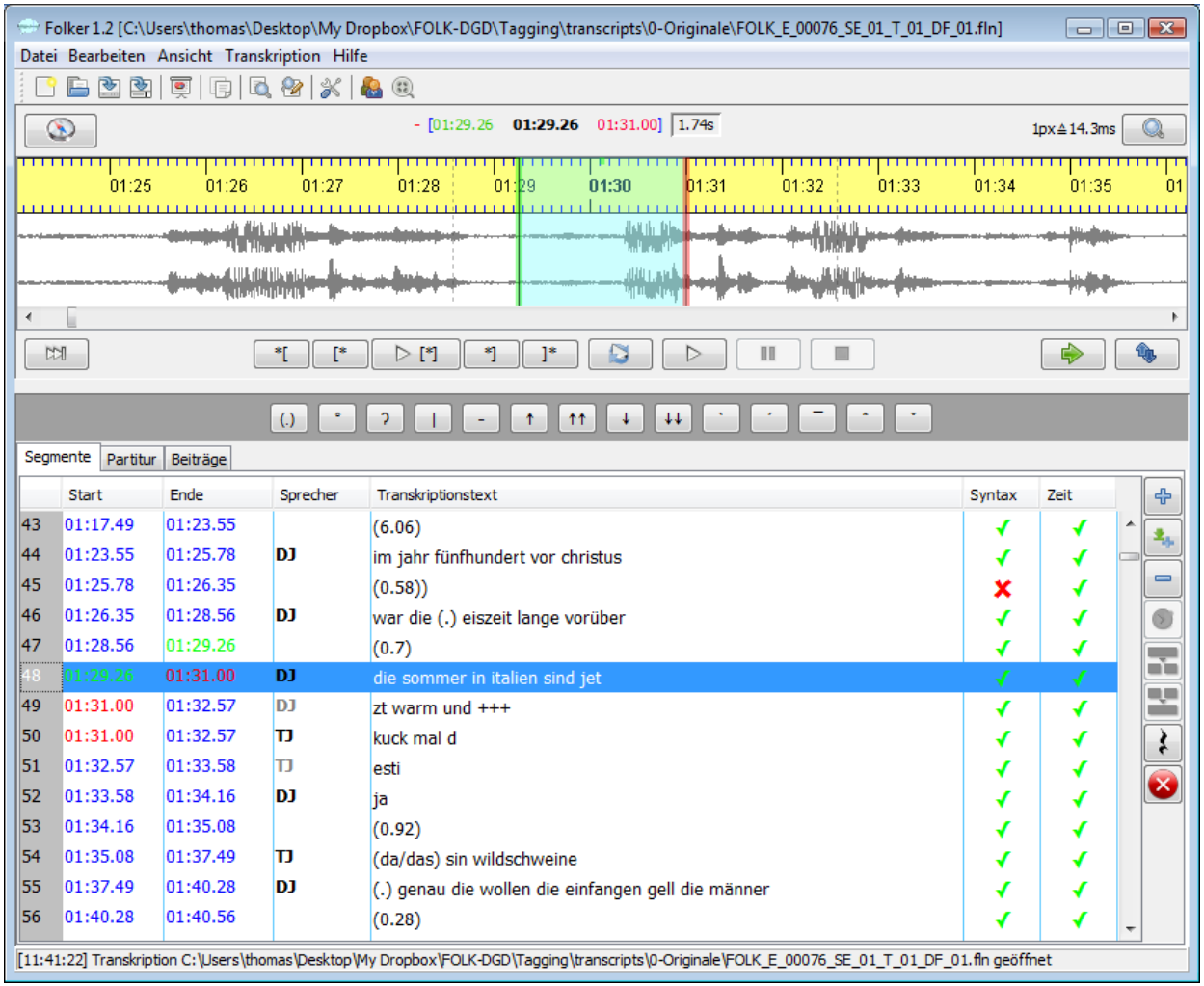
- Semi-automatische Verfahren
  - automatische stochastische und lexikonbasierte Verfahren → Fehlerquoten zwischen 20% und 2%
  - Manuelle Korrektur → Verbessern der stochastischen und lexikonbasierten Verfahren
- in FOLK: eigene Methoden zum Normalisieren, Lemmatisierung und POS-Tagging mit dem TreeTagger nach dem Stuttgart-Tübingen-Tagset, Korrektur im Tool OrthoNormal

# Software-Tools

- Für Transkription und manuelle Annotation
- Anforderungen:
  - Effizienz
  - Präzision, Qualitätskontrolle
  - Interoperabilität
- Anforderungen variieren projektspezifisch

→ in FOLK: für das Projekt optimierte Werkzeuge, basierend auf etablierten Mehrzweckinstrumenten (EXMARaLDA)

# FOLK-EditoR



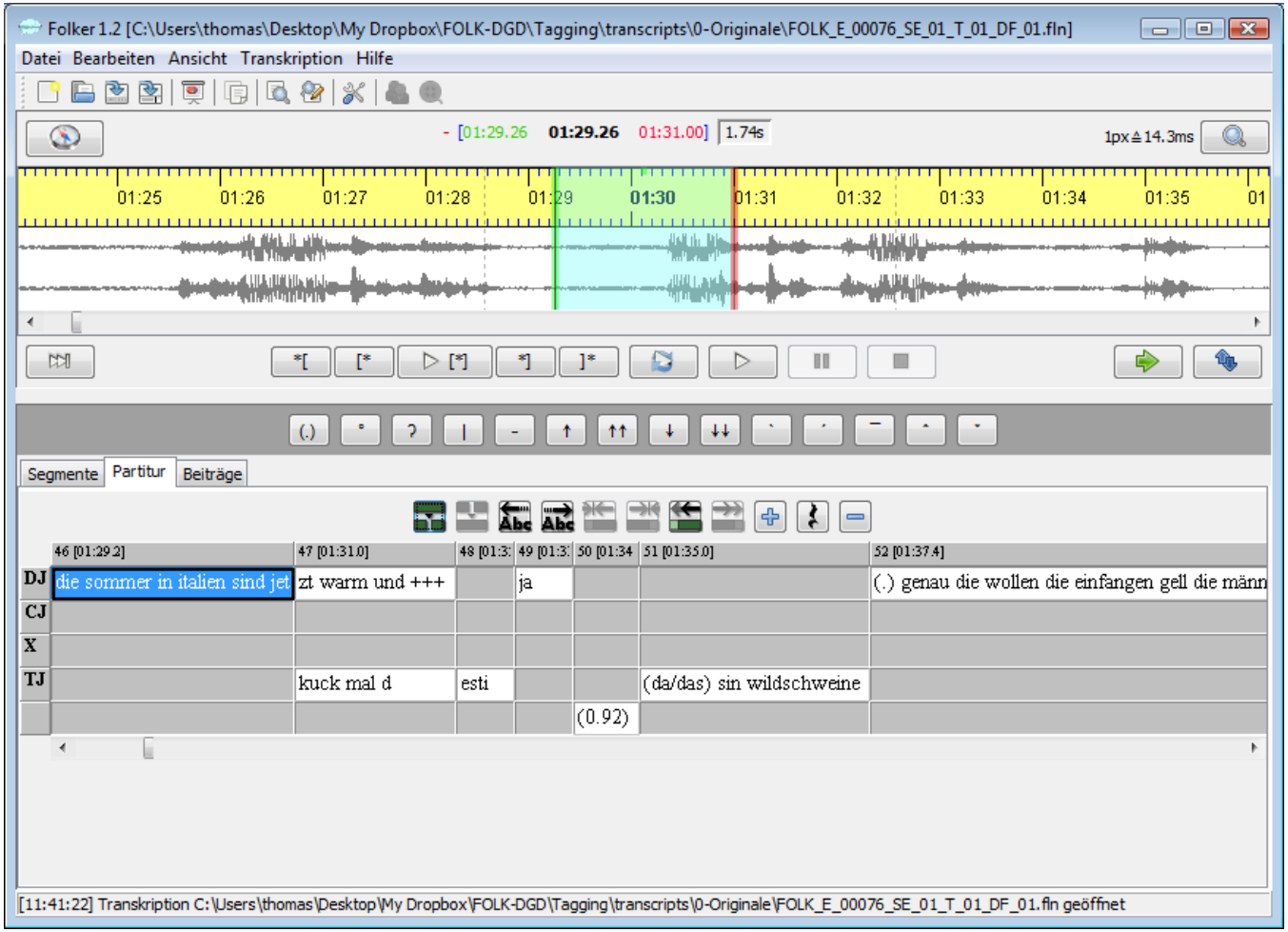


AGD

Thomas Schmidt / Arnulf Deppermann
Archiv für gesprochenes Deutsch

INSTITUT FÜR DEUTSCHE SPRACHE

# FOLK-EditoR

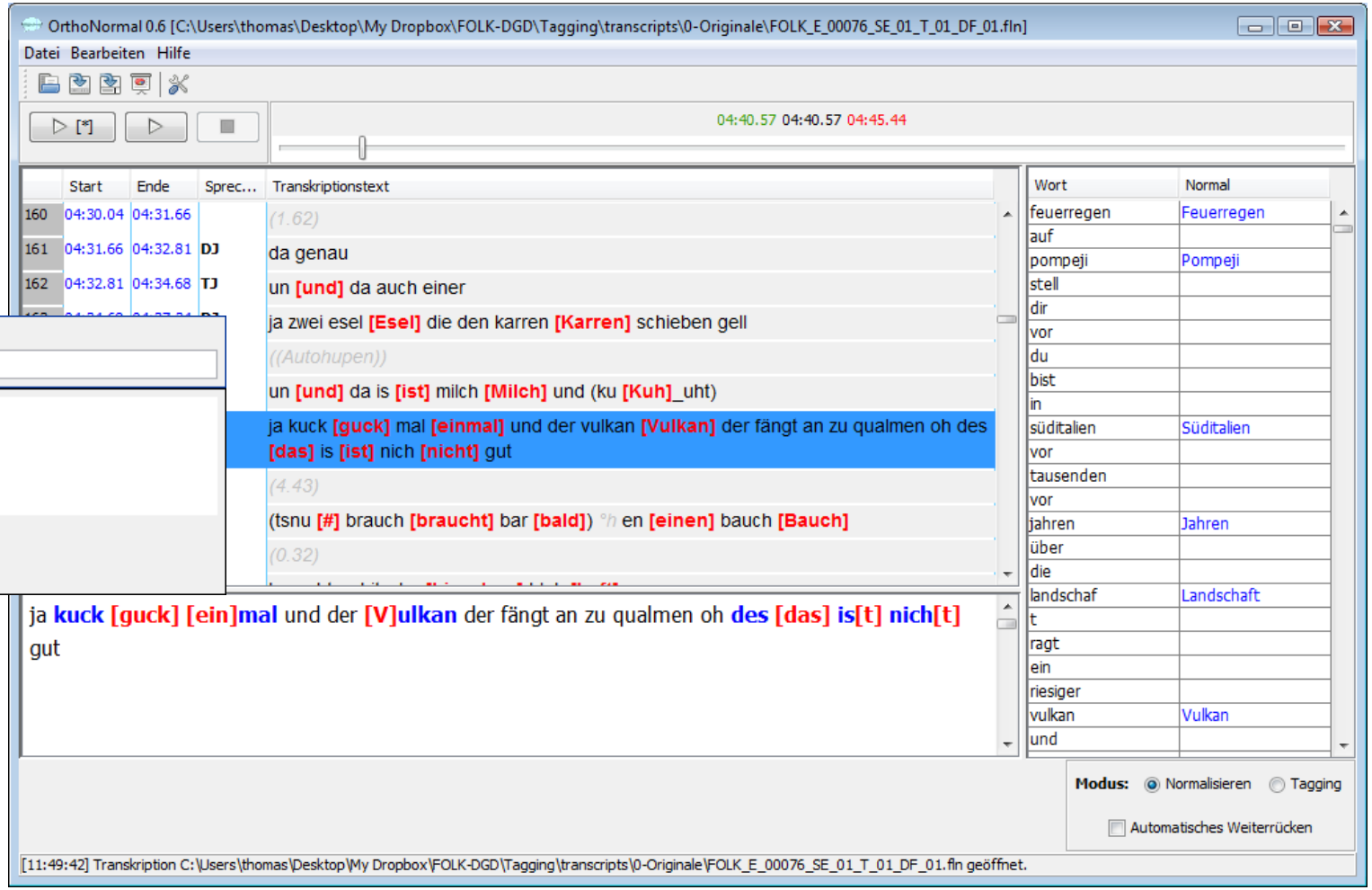




Thomas Schmidt / Arnulf Deppermann
Archiv für gesprochenes Deutsch

INSTITUT FÜR DEUTSCHE SPRACHE

# OrthoNormal



# Rechtliche Aspekte

- Datenschutzrecht erfordert Einverständniserklärung der aufgenommenen Personen (*informed consent*)
- Zusagen:
  - Anonymisierung der Audio-Daten und Metadaten
  - Pseudonymisierung der Transkriptdaten
  - Zugangskontrolle für veröffentlichte Daten (Nutzung nur für Forschung und Lehre, keine Weitergabe)

# Rechtliche Aspekte

- Umsetzung:
  - Informed Consent **vor** der Aufnahme
  - Maskierung (Verrauschen) von Nennungen von Personennamen u.Ä. in den Aufnahmen
  - Maskierung (Ersetzung durch Pseudonyme) von Personennamen u.Ä. in den Aufnahmen
  - „Verrauschen“ der Metadaten (z.B. aus „Oggersheim“ wird „Rheinfränkische Sprachregion“, nur Geburtsjahr des Sprechers)
  - Restriktiver Umgang mit personenbezogenen Daten im Projekt, besonders restriktiv bei sehr sensiblen Daten (z.B. Schichtübergabe im Krankenhaus)
  - Persönliche Prüfung von Anmeldungen bei der DGD
- Zeitaufwändig
- Im großen und ganzen praktikabel
- Andere rechtliche Probleme bei Aufnahmen aus Rundfunk und Fernsehen

# Schulung und Beratung

- Publikation von Verfahren (AV-Aufnahme, Datenschutz, Digitalisierung, Transkription, Anonymisierung etc.), Standards, Tools und Manualen im Gesprächsanalytischen Informationssystem GAIS

  http://prowiki.ids-mannheim.de/bin/view/GAIS/
- Schulungen
- Unterstützung von Forschungsprojekten nach Prinzip „Beratung im Austausch gegen Archiv-Daten“: Sicherung von Standardkonformität und Vollständigkeit

# Perspektiven

- Ausbau der Funktionalität der DGD in den kommenden Jahren (POS-Tagging, Video, komplexe und gesprächsstruktursensitive Suchen, persönliche Arbeitsbereiche …)
- Ausbau von FOLK in den kommenden Jahren
  - Mind. 20h transkribierte Aufnahmen pro Jahr
  - Weitere Datenspenden
  - Workflow für und Integration von **Videodaten**
- Einbindung der DGD in digitale Infrastrukturen

CLARIN-D
Forschungsinfrastruktur für Sprachressourcen in den Geistes- und Sozialwissenschaften

hzsk hamburger zentrum für sprachkorpora

Mehrsprachige (mündliche) Korpora


Dokumentation bedrohter Sprachen

Korpora des gesprochenen und geschriebenen Deutsch

**Bayerisches Archiv für Sprachsignale**

Korpora aus der / für die Sprachtechnologie

# CLARIN-Infrastruktur

- Gemeinsame Metadatendomäne
  - Auffinden von Ressourcen
- Federated Search
  - Gemeinsames Abfragen von Ressourcen an verschiedenen Standorten
- Standardisierte Schnittstellen
  - Zugriff auf (semi-)automatische Verfahren

# Beispiele

- Zweitspracherwerb vs. Erstspracherwerb
- Belege gesprochener Sprache in elektronischen Wörterbüchern
- Geschriebenes vs. gesprochenes Deutsch
- Automatisches Ton-Text-Alignment
- Topic Detection in biographischen Interviews
- Verknüpfung mit digitalen Sprachatlanten
- …

| Speaker | Left Context | Match | Right Context |
|---|---|---|---|
| Dim | und ähm ((1,0s))f/ also bis zum | Ende | natürlich links und ziehst dan |
| Ham | ((1,0s)) nicht bis zum | Ende | du darfst die Linie nicht ((0,6 |
| Lucy | ((0,5s)) und dann ((2,2s))bis | Ende | den/ ((0,5s)) der Nadeln und |
| Hoa | ((0,3s)) bis zum | Ende | bis es nicht mehr möglich ist |
| Mat | ((0,2s)) bis zum | Ende | der/ wo die Wagen sind ((2,9 |
| Mat | du gehst bis zum | Ende | des ((0,5s)) Wagens wo/ der |
| Dav | ((0,6s)) bis man ähm am unteren | Ende | einer Sanduhr angekommen |
| Tan | e dass) ((0,6s)) ähm ((2,4s))bis | Ende | / äh ((0,2s)) ((stockt)) bis End |
| Tan | nde/ äh ((0,2s)) ((stockt)) bis | Ende | des Blattes also |
| Vic | ((0,6s)) | Ende | es Karavan-Bild ((2 |
| Hus | und dann hin/ wenn du so an dem | Ende | des Kartons gekommen bist |
| Ali | ((0,7s)) am | Ende | des Bildes dann äh ((0,9s)) ja |
| Eli | bi/ nach unten genau bis zum | Ende | des Bildes |
| Eli | ((0,3s)) ja ((2,6s)) bis zum | Ende | des Bildes ((4s)) also du geh |
| Eli | (1s)) und deine Linie/ ((0,2s)) | Ende | deine Linie |
| Eli | nn deine/ ähm ((0,1s)) also die | Ende | / ((0,4s)) äh ((unverständlich, |
| Eli | äh ((unverständlich, 0,2s)) das | Ende | deiner Linie ((0,7s)) befindet |
| Eli | mit dem/ mit den Nägel ((2,1s)) | Ende | deiner Linie |
| Eli | | Ende | der Linie sollst du unter/ in d |
| Eli | h links ((0,8s)) aber nicht zum | Ende | des Blattes |

**http://agd.ids-mannheim.de**

**http://dgd.ids-mannheim.de**

AGD

INSTITUT FÜR DEUTSCHE SPRACHE